VIETNAM GENERAL CONFEDERATION OF LABOR
TON DUC THANG UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY

NGUYEN QUANG HUY - 523H0140
NGUYEN GIA NHAT KHANH - 523H0149

MIDTERM ESSAY

DATA ANALYTICS AND VISUALIZATION

HO CHI MINH CITY, 2025

VIETNAM GENERAL CONFEDERATION OF LABOR
TON DUC THANG UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY

NGUYEN QUANG HUY - 523H0140
NGUYEN GIA NHAT KHANH - 523H0149

MIDTERM ESSAY

DATA ANALYTICS AND VISUALIZATION

Advised by
Dr. Tran Luong Quoc Dai

HO CHI MINH CITY, 2025

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Dr. Tran Luong Quoc Dai, our instructor and mentor, for his valuable guidance and support throughout the mid-term report. He has been very helpful and patient in providing us with constructive feedback and suggestions to improve our work. He has also encouraged us to explore new technologies and techniques to enhance our system's functionality and performance. We are honored and privileged to have him as our teacher and supervisor.

Ho Chi Minh city, 9nd November 2025.

Author

(Signature and full name)

Huy

Nguyen Quang Huy

Khanh

Nguyen Gia Nhat Khanh

DECLARATION OF AUTHORSHIP

We hereby declare that this is our own project and is guided by Dr. Tran Luong Quoc Dai. The content research and results contained herein are central and have not been published in any form before. The data in the tables for analysis, comments and evaluation are collected by the main author from different sources, which are clearly stated in the reference section.

In addition, the project also uses some comments, assessments as well as data of other authors, other organizations with citations and annotated sources.

If something wrong happens, we'll take full responsibility for the content of my project. Ton Duc Thang University is not related to the infringing rights, the copyrights that We give during the implementation process (if any).

Ho Chi Minh city, 9nd November 2025.

Author

(Signature and full name)

Huy

Nguyen Quang Huy

Khanh

Nguyen Gia Nhat Khanh

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Financial markets are complex, dynamic systems driven by a vast number of factors, from firm-specific performance to global macroeconomic shifts. Understanding the behavior of these markets requires data analysis to uncover patterns, manage risk and make decisions.

The primary objective of this report is to conduct an analysis of the **S&P 500 Index** (representing the performance of 500 of the largest U.S. companies) by applying a structured data analysis methodology. This report will conduct statistical and visual techniques to analyze the index's historical behavior.

To achieve this, the analysis will utilize a comprehensive time-series dataset of the S&P 500 Index merged with macroeconomic indicators (CPI, GLD, IEF, VIXY, UUP, USO). The final dataset spans from **January 4, 2011**, to **December 12, 2023**.

This report is structured into five core sections as required by the project syllabus:

1. **Exploratory Data Analysis (EDA):** To summarize, visualize, and identify the main characteristics of the data.
2. **Probability Distribution Analysis:** To test whether the index's returns follow a standard normal distribution.
3. **Hypothesis Testing:** To apply statistical tests to answer a specific research question about market behavior.
4. **Correlation Analysis:** To quantify the relationships between the financial index and key macroeconomic variables.
5. **Multiple Linear Regression:** To build and evaluate a model aimed at explaining the drivers of the index's returns.

CHAPTER 1. EXPLORATORY DATA ANALYSIS

1.1. Data Summary

The dataset was collected from two main sources: Yahoo Finance (yFinance) for daily index values and Alpha Vantage for macroeconomic indicators. The objective of this analysis is to perform Exploratory Data Analysis (EDA) to understand the dataset's structure, trends, and potential relationships among variables.

- Number of records (rows): 3257
- From 4th January, 2011 to 12th December, 2023 (04/01/2011 - 12/12/2023)
- Number of features (variables): 11

Variable	**Type**	**Description**
Date	datetime	The trading date corresponding to each record
Open	float	Opening price of the index on a given trading day
High	float	Highest price of the index during the trading day
Low	float	Lowest price of the index during the trading day
Close	float	Closing price of the index at the end of the trading day
Volume	int	The total number of shares traded for the index constituents on that day
VIXY	float	CBOE Volatility Index, representing expected market volatility
IEF	float	U.S. 10-Year Treasury Yield, reflecting long-term interest rate expectations
UUP	float	U.S. Dollar Index, measuring the value of the U.S. dollar against a basket of major currencies
USO	float	Crude oil price per barrel (WTI benchmark)
GLD	float	Gold spot price per ounce, representing a key safe-haven asset
CPI	float	U.S. Consumer Price Index, representing inflation levels

Table 1.1.1. Data Type and Description

1.2. Missing and Duplicated Data Processing

1.3. Descriptive Statistical Analysis

	mean	median	std	min	max	q1	q3	iqr
Open	2.634.749	2429.2	1.029.353	1097.42	4804.51	1861.46	3380.86	1519.4
High	2.649.095	2439.27	1.035.696	1125.12	4818.62	1873.34	3399.54	1526.2
Low	2.619.422	2413.54	1.022.738	1074.77	4780.04	1849.69	3366.15	1516.46
Close	2.635.274	2429.01	1029.48	1099.23	4796.56	1862.49	3383.54	1521.05
Volume	3.890.499.324.532	3735980000.0	928.314.382.463	0.0	9976520000.0	3336880000.0	4253920000.0	9170400 0.0
CPI	26.661	26.67	0.86	23.92	28.52	26.108	27.399	1.29
GLD	142.763	137.97	24.546	100.5	193.89	120.78	166.39	45.61
IEF	106.222	105.8	6.886	88.95	123.06	102.14	109.99	7.85
USO	31.879	32.12	20.762	2.13	91.99	12.28	39.39	27.11
UUP	24.727	24.93	2.226	20.93	30.67	22.548	26.07	3.523
VIXY	25.921	21.24	17.384	5.25	121.95	14.62	29.97	15.35

Table 1.3.2. Data Descriptive Statistical

1.3.1. SP500

- Prices fluctuate widely over time; large std and range.
- Mean and median are close, suggesting a roughly symmetric distribution, though extreme highs exist.
- Volume shows large variability and may need outlier handling before analysis.

Column	Analysis
Open	The mean and median are close so the distribution is roughly symmetric. Large standard deviation indicates high price variability over the period.
High	Similar to Open; the daily high is slightly above Open/Close.
Low	Daily low is lower than Open/High; variability is high.
Close	Close price is close to Open mean → intraday movements roughly balanced.
Volume	Extremely variable; zero volume on some days may indicate missing data or

	non-trading days.

Table 1.3.3. SP500 Index Descriptive Statistical Analysis

1.3.2. Macroeconomic indicators

- CPI, IEF, and UUP are stable → suitable as benchmarks or for long-term correlations.
- GLD, USO, VIXY are highly variable → reflect market risk, commodity price swings, or investor sentiment.
- Small IQR → data concentrated; large IQR → wider distribution, more prone to outliers.

Column	**Analysis**
CPI	Inflation is stable; very low variability.
GLD	Gold prices show moderate variability; large max-min reflects market spikes during crises.
IEF	Medium-term government bonds are stable; low variability.
USO	Oil prices are highly volatile; large std indicates significant market swings.
UUP	The USD index is relatively stable; low variability.
VIXY	Volatility index shows strong fluctuations; extreme spikes exist (fear spikes).

Table 1.3.4. Macroeconomics Indicators Descriptive Statistical Analysis

1.3.3. Comparison and Insights

- SP500 price columns (Open, High, Low, Close) are highly correlated and show large variability.
- Volume is extremely variable → consider log transformation or normalization for regression/correlation analysis.
- Stable economic indicators (CPI, IEF, UUP) are useful for long-term trend analysis.

- Volatile indicators (USO, GLD, VIXY) should be normalized (e.g., Min-Max scaling) when plotting together with SP500 to compare relative movements.

1.4. Data Visualization

Due to the characteristics of the stock market dataset, the Close price is often considered the most representative indicator of daily performance, as it reflects the final consensus of value after intraday fluctuations. Therefore, subsequent analyses primarily focus on the Close price, along with selected macroeconomic indicators and technical stock indicators, while Open, High, and Low are excluded from detailed visualization (except Candlestick Chart).

1.4.1. Candlestick Chart + Volume Bar Chart

- Apply to: Open, High, Low, Close Price and Volume.
- Purpose: This pair effectively visualizes price movement patterns and market strength through trading volume.
- Why chosen: Candlesticks reveal intra-period volatility, while volume bars explain how strongly each movement is supported by market activity.

Figure 1.4.1. Candlestick chart for Open, High, Low, Close Price and Trading Volume

1.4.2. Boxplot

- Apply to: Representative numerical features such as technical indicators or macroeconomic factors.
- Purpose: Shows data spread, quartiles, and outliers.
- Why chosen: Directly supports the Anomaly Detection part (1.5) and helps compare scale and variability between variables.
- Normalization: All selected numerical features are scaled to a [0, 1] range using Min-Max normalization to remove differences in units and magnitudes. This ensures that each variable contributes proportionally when visualizing boxplots, allowing clearer comparison of distributions, quartiles, and outliers across diverse indicators.

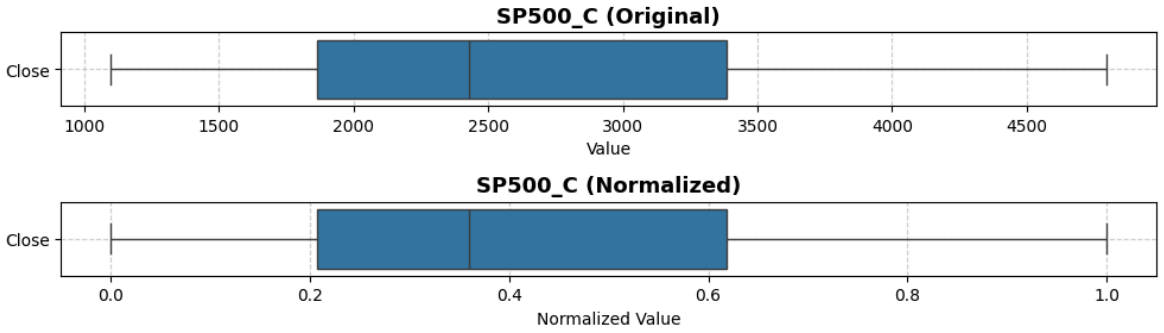

Figure 1.4.2. Boxplot of SP500 Close Index

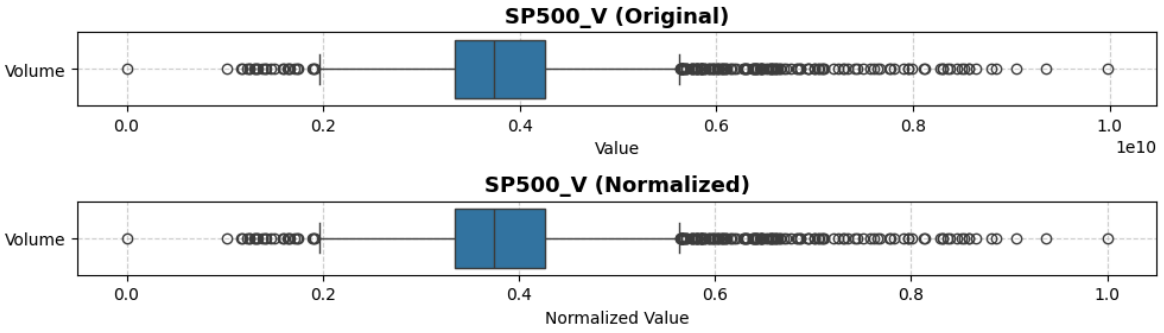

Figure 1.4.3. Boxplot of SP500 Trading Volume Index

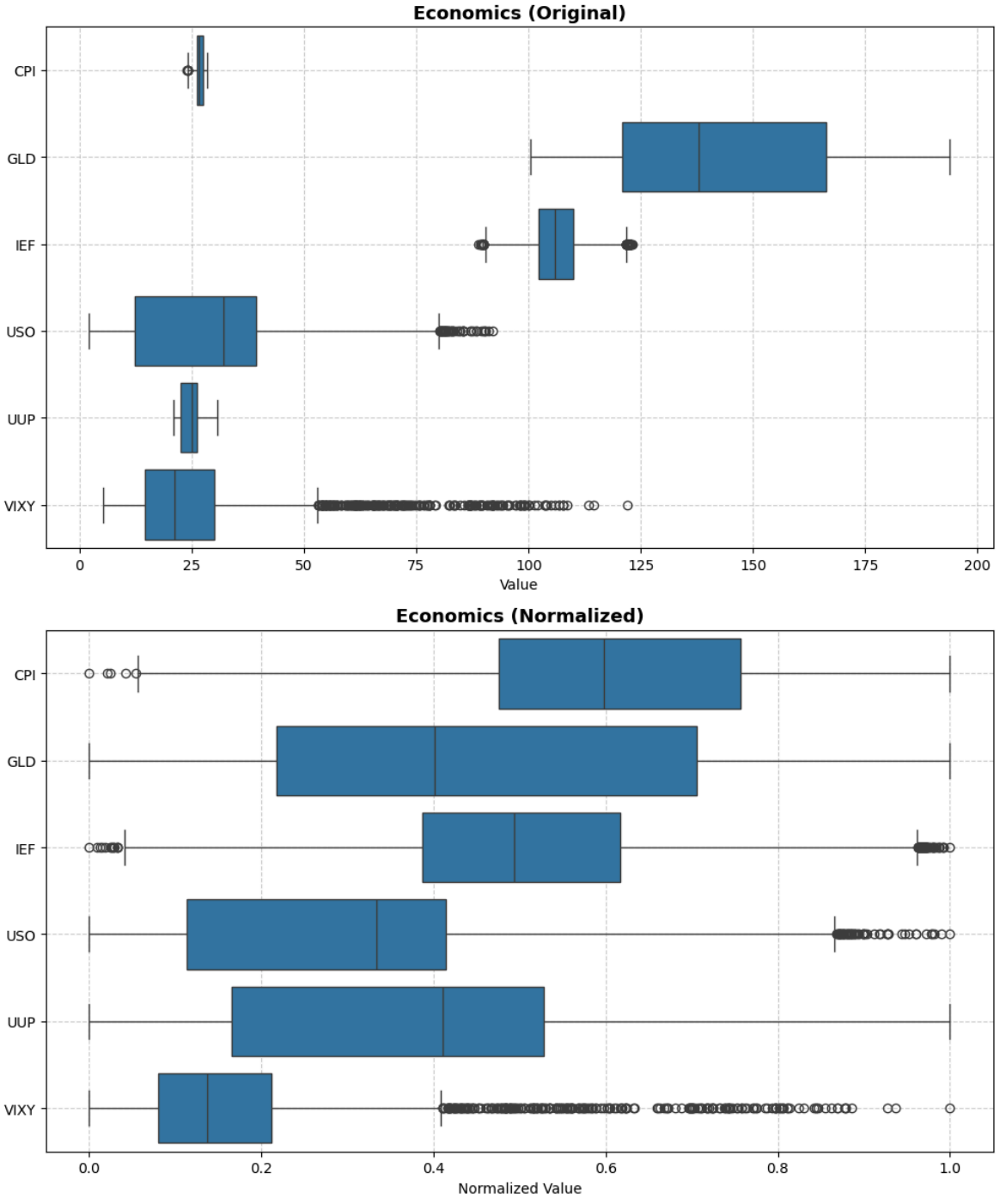

Figure 1.4.4. Boxplot of Macroeconomics Indicators

1.4.3. Histogram

- Apply to: Representative numerical features such as SP500 Close, Trading Volume, and macroeconomic indicators (CPI, GLD, IEF, USO, UUP, VIXY).

- Purpose: Visualizes the distribution of individual variables, showing frequency, skewness, and potential multimodal patterns.
- Why chosen: Histograms provide insight into how data is distributed across values, highlighting asymmetries, peaks, and potential anomalies that may not be visible in summary statistics alone.
- Normalization: When comparing multiple variables with different units, values are scaled to a [0, 1] range using Min-Max normalization to allow consistent visual comparison of distributions across variables.

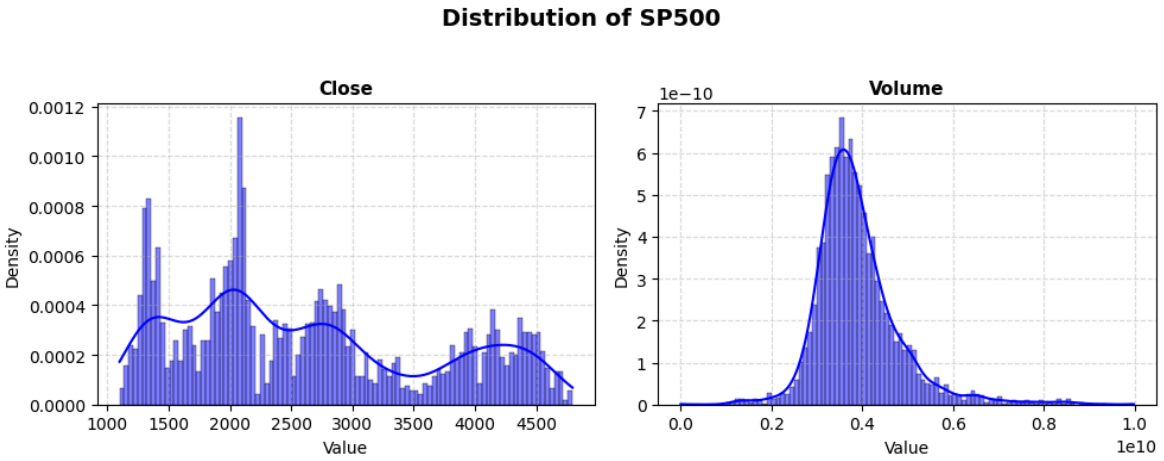

Figure 1.4.5. Histogram of SP500 Close Index and Trading Volume

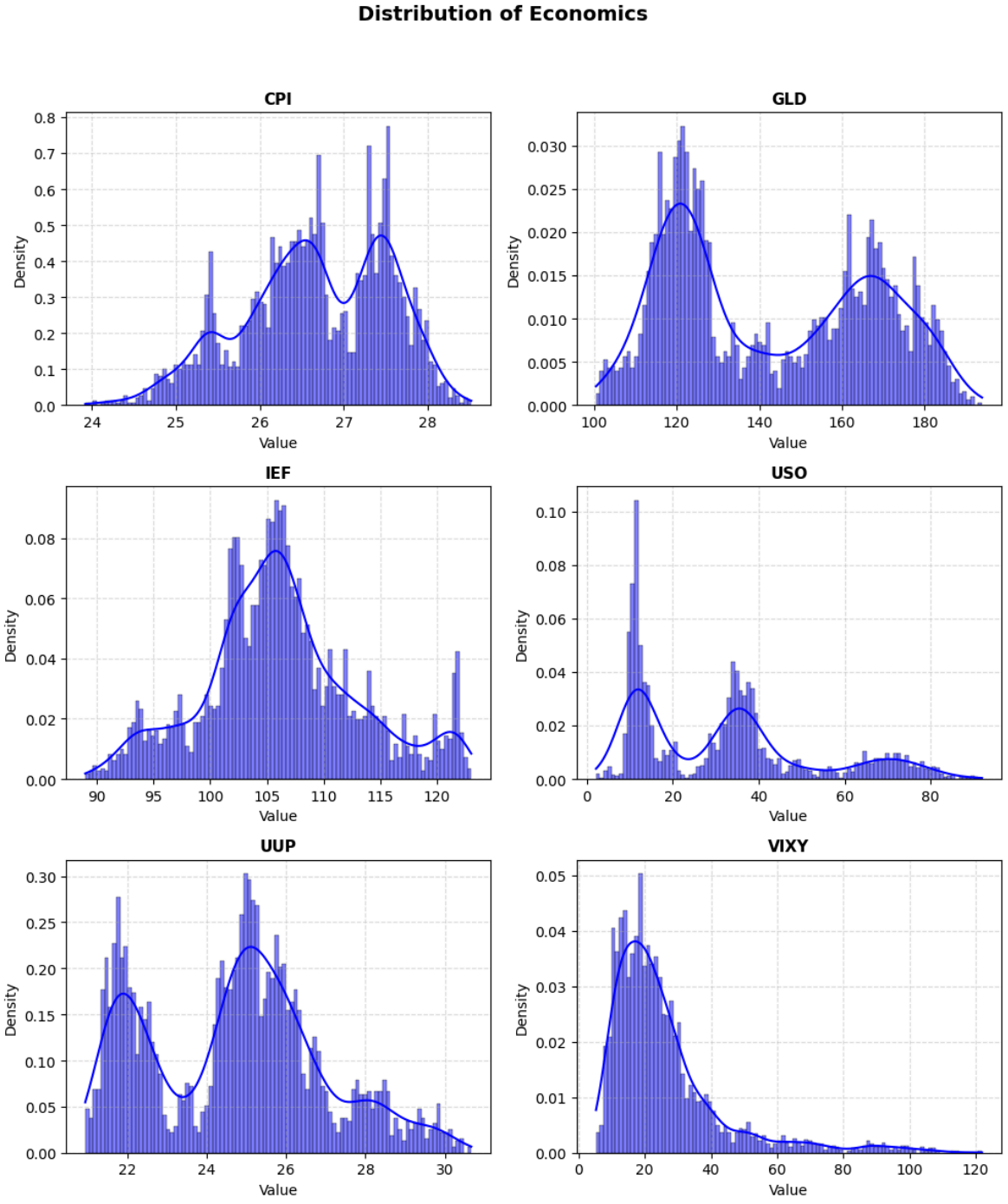

Figure 1.4.6. Histogram of Macroeconomics Indicators

1.4.4. Line Chart

- Apply to: SP500 Close index, Trading Volume, and macroeconomic indicators over time.

- Purpose: Shows temporal trends, patterns, and potential correlations between variables.
- Why chosen: Line charts effectively illustrate how values evolve over time, making it easier to identify trends, seasonality, sudden spikes, and relationships between market performance and macroeconomic factors.
- Normalization: To compare multiple indicators on the same chart, all selected numerical features are scaled to a [0, 1] range using Min-Max normalization, ensuring that each variable contributes proportionally and trends are visually comparable.

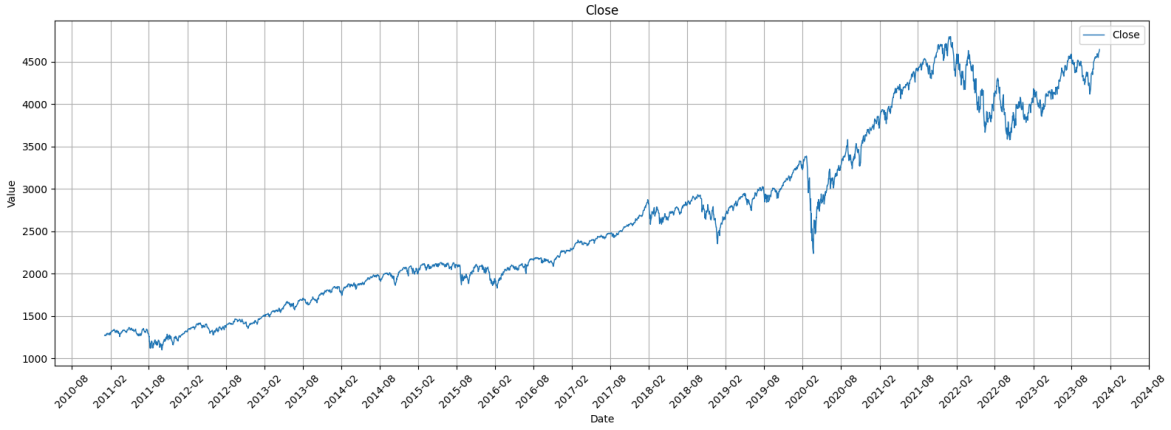

Figure 1.4.7. Line Chart of SP500 Close Index

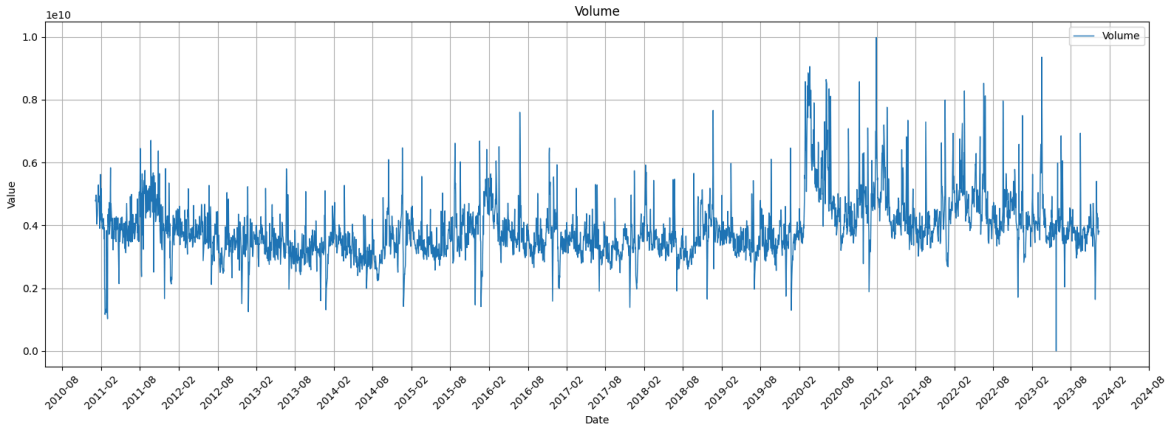

Figure 1.4.8. Line Chart of SP500 Trading Volume

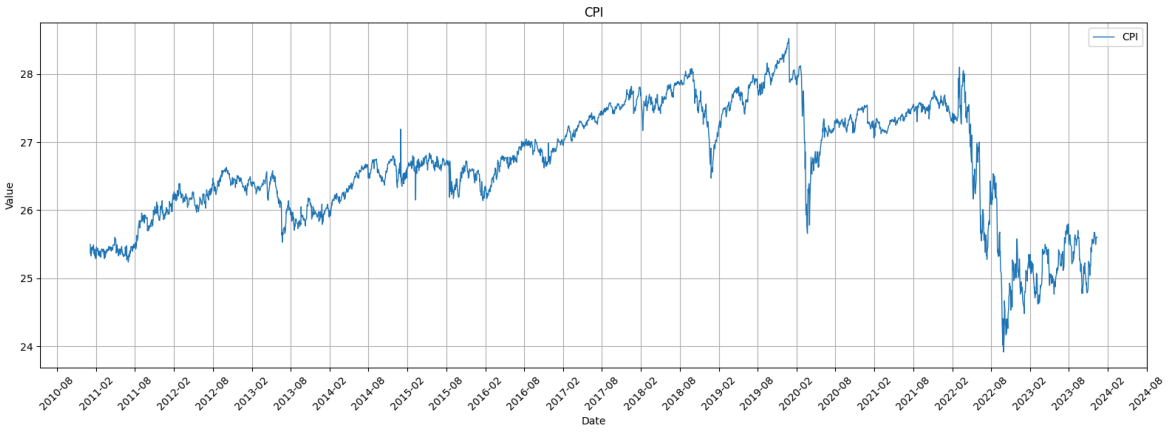

Figure 1.4.9. Line Chart of CPI

Figure 1.4.10. Line Chart of GLD

Figure 1.4.11. Line Chart of IEF

Figure 1.4.12. Line Chart of USO

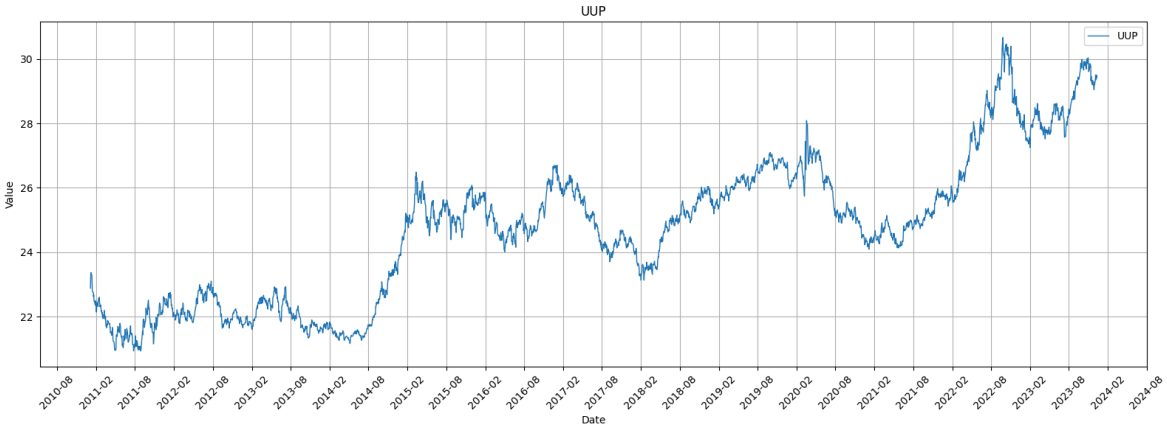

Figure 1.4.13. Line Chart of UUP

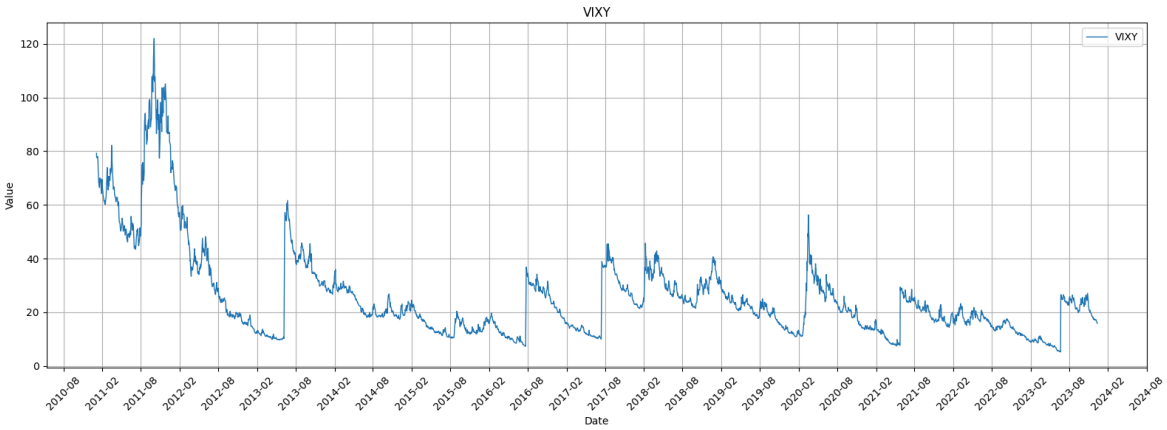

Figure 1.4.14. Line Chart of VIXY

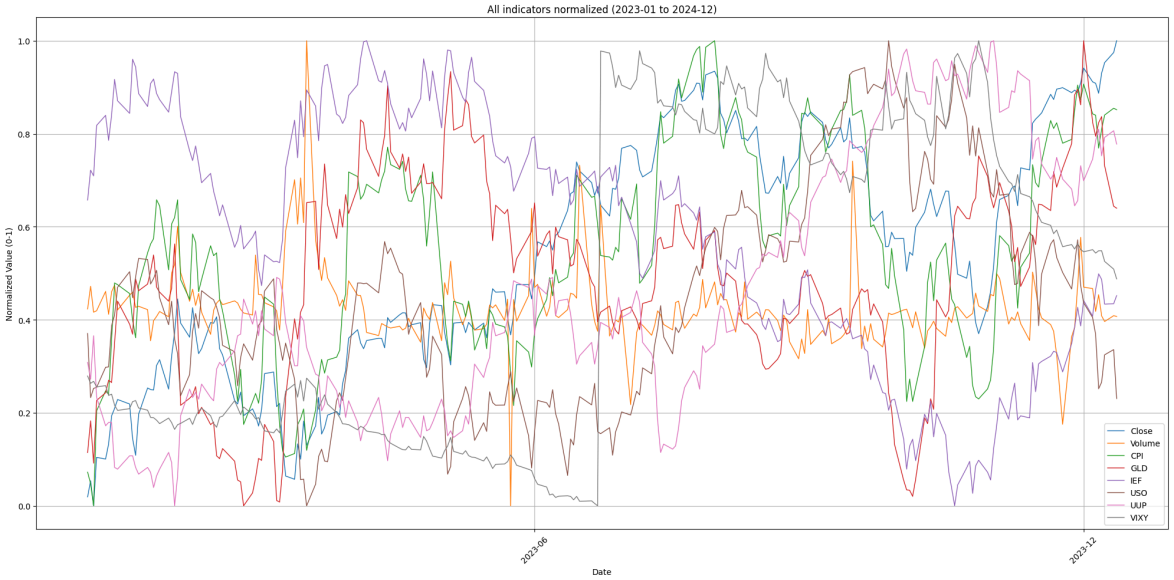

Figure 1.4.15. Line Chart of All Columns (2023-2024)

1.4.5. Heatmap

- Apply to: Correlation matrix of SP500 Close, Trading Volume, and macroeconomic/technical indicators.
- Purpose: Visualizes the strength and direction of relationships among variables, helping identify positive or negative correlations.
- Why chosen: Heatmaps provide an intuitive way to detect strong associations or dependencies between variables, which can guide feature selection, multivariate analysis, and anomaly detection.
- Normalization: Correlation values are inherently bounded between -1 and 1, so no additional normalization is required. Color intensity directly represents correlation magnitude, allowing immediate visual assessment of inter-variable relationships.

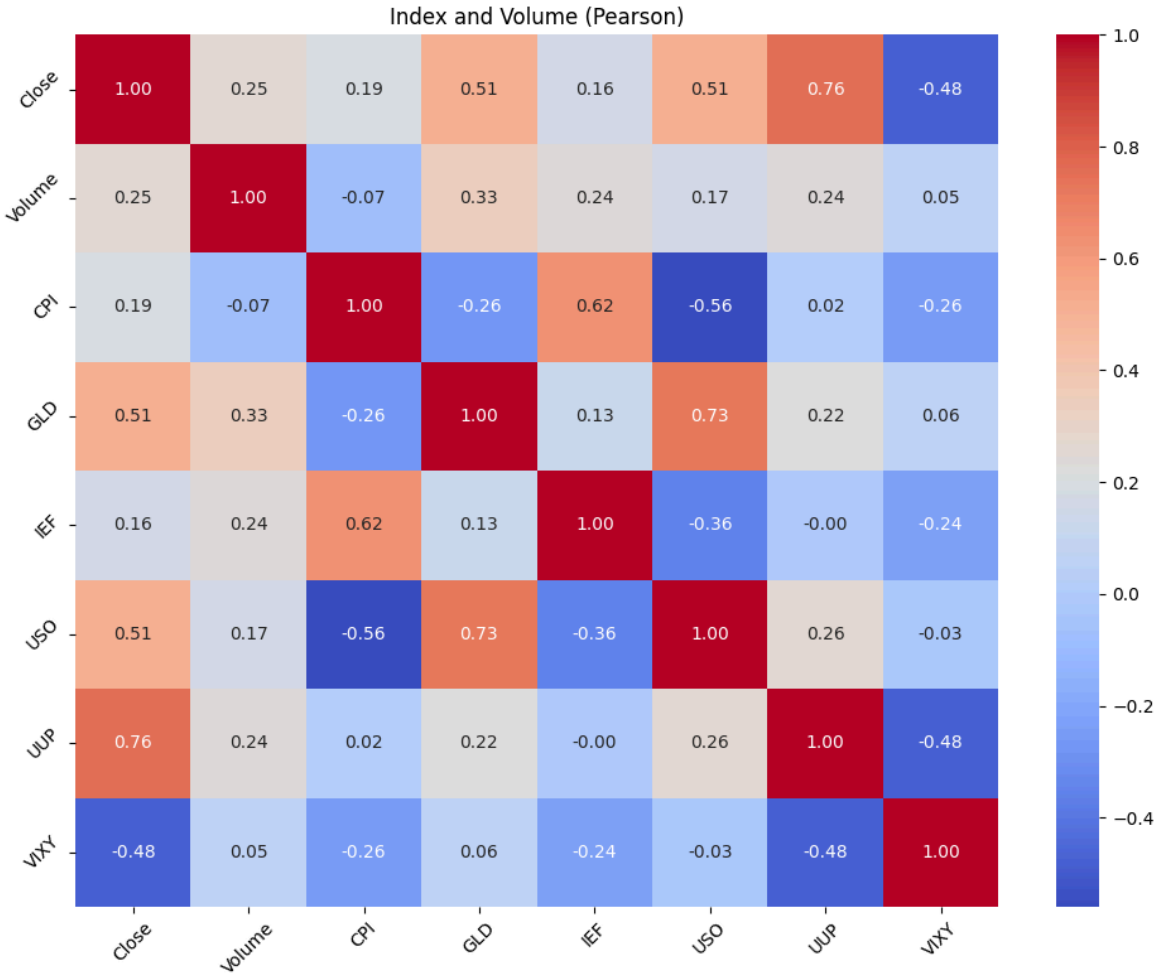

Figure 1.4.16. Correlation matrix with Pearson method

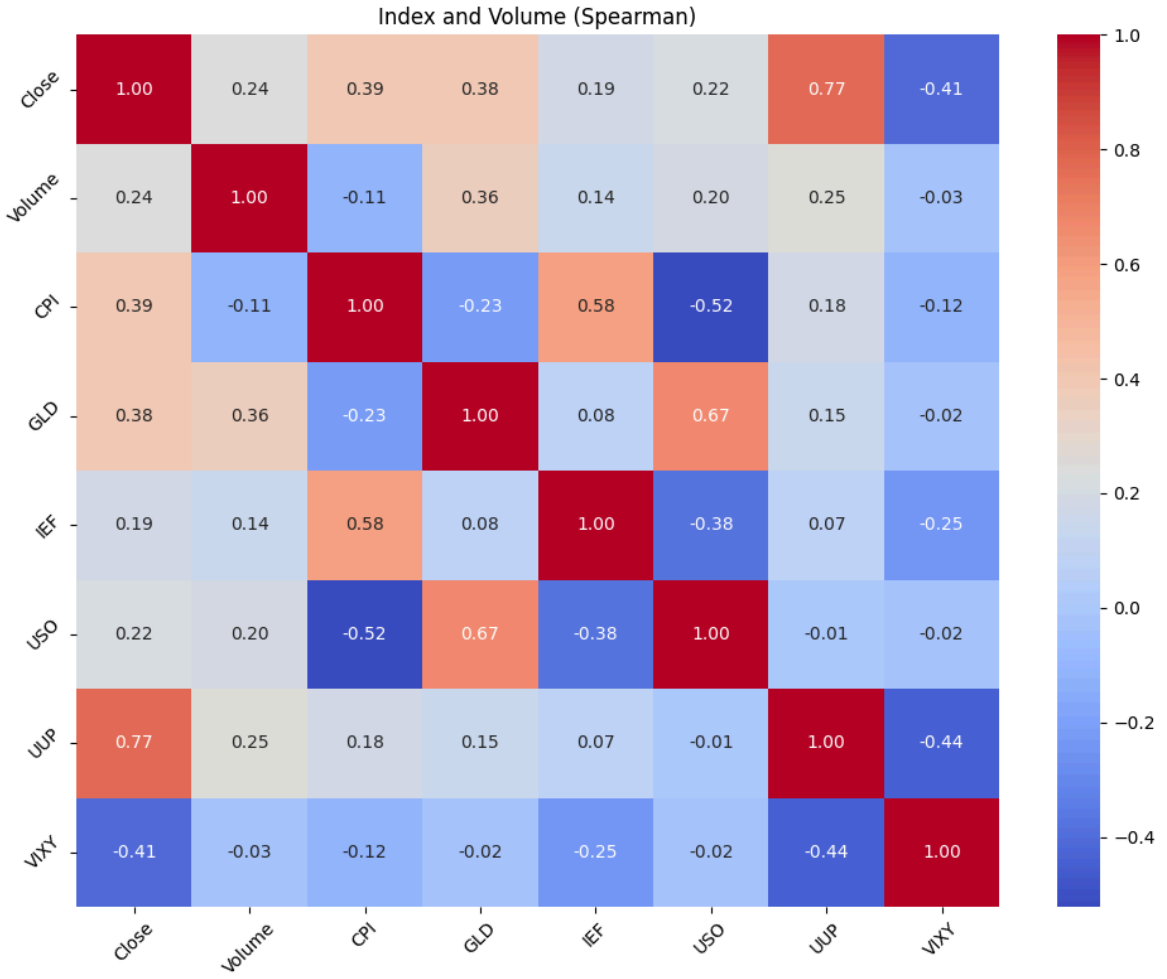

Figure 1.4.17. Correlation matrix with Spearman method

1.5. Anomaly Detection

To identify anomalies in the dataset, multiple methods were applied to both the SP500 Close price and other numerical indicators. The methods include:

1.5.1. Methodology

a. Z-score Method

- Measures the number of standard deviations a value is from the mean.
- Threshold of 3.0 was used; values beyond ±3 standard deviations were flagged as outliers.

- Applied to all numerical features in the dataset.

b. Interquartile Range (IQR) Method

- Calculates the difference between the 75th percentile (Q3) and 25th percentile (Q1).
- Observations outside [Q1 - 1.5×IQR, Q3 + 1.5×IQR] are considered outliers.
- Captures extreme values in skewed distributions.

c. Rolling Z-score

- For time series, computes z-score based on a rolling window of 20 periods.
- Flags observations that deviate significantly from their local context, capturing short-term spikes.

d. Seasonal Decomposition Residuals

- Decomposes the time series into trend, seasonal, and residual components.
- Outliers are identified from residuals exceeding 3 standard deviations.
- Highlights points deviating from expected seasonal patterns.

e. ARIMA Residuals

- Fits an ARIMA(1,1,1) model to the Close price and examines residuals.
- Residuals with absolute z-score above 3.5 are flagged.
- Captures anomalies unexplained by autoregressive trends.

f. Isolation Forest

- A machine learning approach detecting anomalies in multivariate space.
- Applied to all numerical features after standard scaling.

- Flags observations that are isolated in feature space, complementing univariate methods.

1.5.2. Aggregation of Outliers

Out of the full dataset, multiple significant spikes in the Close price were detected by several methods.

Points flagged by multiple methods tend to coincide with periods of extreme market volatility, consistent with known historical events (e.g., market crashes or rallies).

The plot_time_series_with_flags visualization highlights these points:

- Red dots mark outliers, with size and opacity proportional to the number of methods flagging them.
- This allows easy identification of strong anomalies versus minor deviations.

Date	2020-03-26	2020-03-27	2020-03-28	2020-03-28
z_close	True	True	False	False
iqr_close	False	False	True	False
decomp_close	False	True	False	False
arima_close	False	False	False	False
iso_multivariate	False	True	True	False
rolling_close	False	False	True	False
any_outlier	True	True	True	False
n_methods_flagged	1	3	3	0

Table 1.5.1. Example of Anomalies Flagging

Figure 1.5.1. Anomalies Detection Visualization

CHAPTER 2. PROABILITY DISTRIBUTION ANALYSIS

2.1. Purpose

This step aims to identify the probability distributions of the dataset's quantitative variables. Understanding their shapes helps evaluate market characteristics such as volatility, skewness, and the presence of extreme values. The variables selected represent both stock-specific behavior (Close, Volume) and macroeconomic indicators (IEF, VIXY).

2.2. Visualization

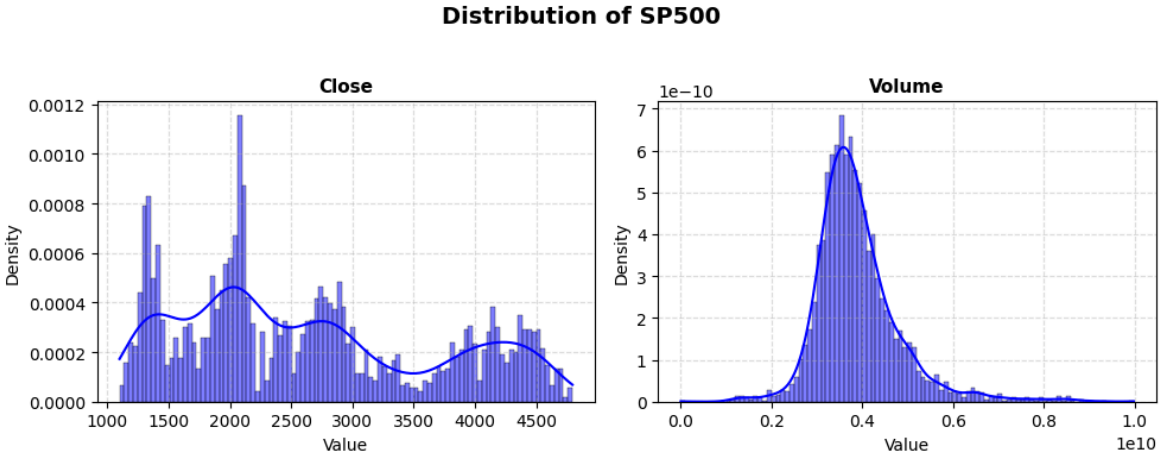

Figure 2.1.1. Distribution of SP500 Close Index and Trading Volume

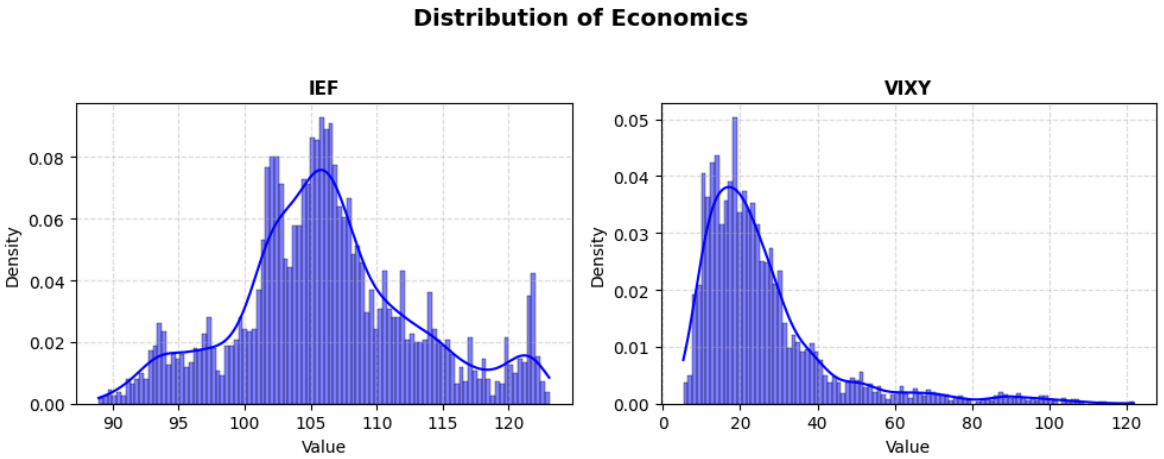

Figure 2.1.2. Distribution of IEF and VIXY

2.3. Analysis

2.3.1. Close

The closing prices show a multimodal distribution with several local peaks around different value ranges (approximately near 1500, 2000, 3000 and 4000). This shape reflects the long historical span of the dataset (2011–2023), during which the S&P 500 underwent multiple growth phases and corrections. The lack of symmetry and presence of multiple peaks indicate that the data does not follow a normal distribution, but rather a composite of several market regimes.

2.3.2. Volume

The trading volume follows a right-skewed (positively skewed) distribution with a sharp peak near 3-4 × 10^9 and a long right tail. This suggests that most trading days maintain moderate liquidity, but occasional spikes in trading activity occur – usually during macroeconomic announcements or crisis events. The right tail indicates the presence of outlier events with exceptionally high trading volume.

2.3.3. IEF

This variable shows a nearly symmetric bell-shaped distribution, though slightly left-skewed. The concentration of values around 105–110 suggests stability typical of fixed-income instruments. The shape is close to a normal distribution, implying that bond ETF prices fluctuate around a consistent mean with low variance – a hallmark of a low-volatility asset.

2.3.4. VIXY

The VIXY distribution is highly positively skewed and leptokurtic (heavy-tailed). Most observations lie in the low range (below 30), but a small number of extreme values stretch up to around 100, reflecting occasional market panic spikes in volatility. This behavior is typical of volatility indices: long periods of calm followed by rare, intense surges.

2.4. Normality Test and Distribution Fit

To verify whether the selected variables follow a normal probability distribution, both the Shapiro–Wilk and Kolmogorov–Smirnov (K–S) tests were applied.

The null hypothesis H0: the variable following a normal distribution was tested at the 5% significance level.

Index	Variable	Shapiro_p	KS_p	Distribution
0	Close	4.426160e-37	3.073666e-43	Non-normal
1	Volume	2.883589e-42	2.667348e-31	Non-normal
2	IEF	2.947230e-21	5.636244e-14	Non-normal
3	VIXY	1.150930e-54	5.995955e-74	Non-normal

Table 2.4.1. Shapiro-Wilk test and Kolomogorov-Smirnov (Ks) test

Since all p-values are far below 0.05, we reject H0 for every variable. Thus, none of the selected variables follow a normal distribution. The strong deviations observed in the histograms and KDE plots (multimodal, right-skewed, and heavy-tailed shapes) are consistent with these results.

These findings confirm that the dataset exhibits non-Gaussian behavior, typical of financial time series, characterized by volatility clustering and extreme fluctuations. For later analysis (hypothesis testing and regression), data transformation (e.g., log-return computation or scaling) may be necessary to improve statistical validity.

CHAPTER 3. HYPOTHESIS TESTING

3.1 Purpose

The goal of this section is to statistically verify whether there are significant differences between different market conditions or economic periods. By performing hypothesis tests on selected quantitative variables, we can determine if the observed changes are statistically meaningful rather than random fluctuations.

3.2. Independent t-test

3.2.1. Research Question

Is there a significant difference in trading volume (Volume) between periods of low volatility and high volatility in the stock market?

3.2.2. Hypotheses

- Null hypothesis (H0): The mean trading volume during low-volatility and high-volatility periods are equal.
- Alternative hypothesis (H1): The mean trading volumes differ between the two periods.

3.2.3. Method

The variable VIXY (Volatility Index ETF) was used as a proxy for market volatility. The dataset was divided into two groups based on the median of VIXY:

- Low-volatility period: VIXY < median(VIXY)
- High-volatility period: VIXY ≥ median(VIXY)

The Independent Two-Sample t-test (assuming unequal variances) was applied to compare the means of Volume between the two groups.

3.2.4. Results

Statistic	**Value**
t-statistic	2.325
p-value	0.02012

Table 3.2.1. T-test result

Since the p-value (0.02012) is less than the significance level $\alpha = 0.05$, we reject the null hypothesis. There is a statistically significant difference in trading volume between low- and high-volatility periods.

During periods of high volatility (high VIXY), the average trading volume tends to increase significantly. This reflects investors' heightened trading activity in

response to market uncertainty – consistent with the behavioral finance notion that fear and uncertainty drive liquidity spikes.

3.2. One-way ANOVA Test

3.3.1. Research Question

Is there a significant difference in the S&P 500 closing price (Close) among different economic periods?

3.3.2. Hypotheses

- Null hypothesis (H0): The mean closing prices are the same across all economic periods.
- Alternative hypothesis (H1): At least one period has a different mean closing price.

3.3.3. Method

The data was divided into three economic phases:

1. Before COVID-19: 2011–2019
2. During COVID-19: 2020–2021
3. After COVID-19: 2022–2023

A One-Way ANOVA (Analysis of Variance) test was conducted to determine whether the average S&P 500 closing prices differ significantly across these periods.

3.3.4. Results

Statistic	**Value**
f-statistic	4588.007
p-value	0.000

Table 3.3.1. ANOVA test result

The p-value is effectively zero ($p < 0.001$), which strongly rejects the null hypothesis. Therefore, there are significant differences in the mean closing price among three economic periods.

The results confirm that the S&P 500 underwent major structural shifts between the pre-COVID, COVID, and post-COVID eras. The sharp increase in mean closing prices after 2020 reflects unprecedented monetary expansion and post-pandemic recovery effects. This demonstrates that macroeconomic regimes have a profound and statistically measurable impact on equity valuations.

CHAPTER 4. CORRELATION ANALYSIS

4.1. Theoretical Basis

- Pearson Correlation: Measures the strength and direction of a linear relationship between two continuous variables. It is sensitive to outliers.
- Spearman Correlation: Measures the strength and direction of a monotonic relationship (i.e., as one variable increases, the other consistently increases or decreases, but not necessarily at a constant rate). It is "non-parametric" and robust to outliers.

4.2. Visualization and Conclusion

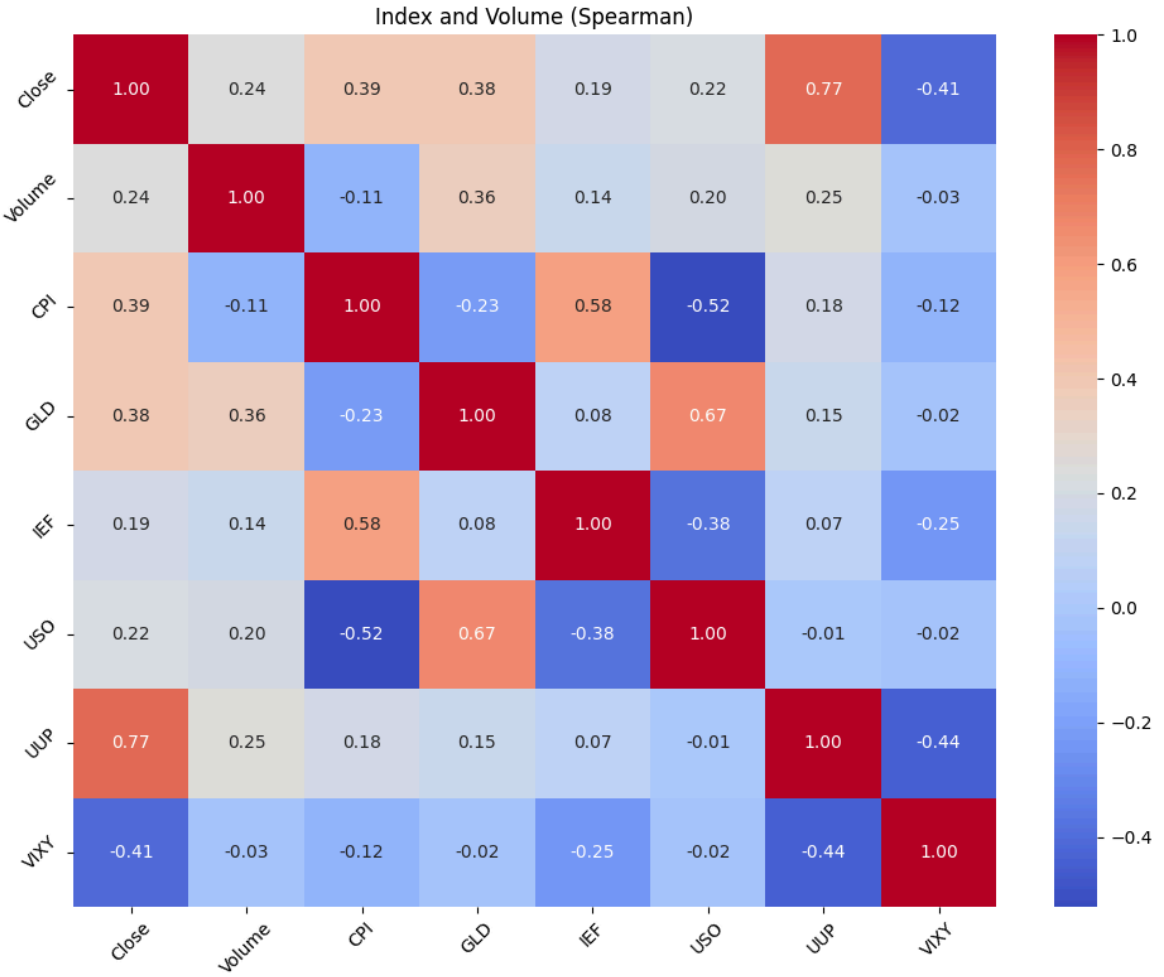

Figure 4.2.1. Spearman Correlation Matrix

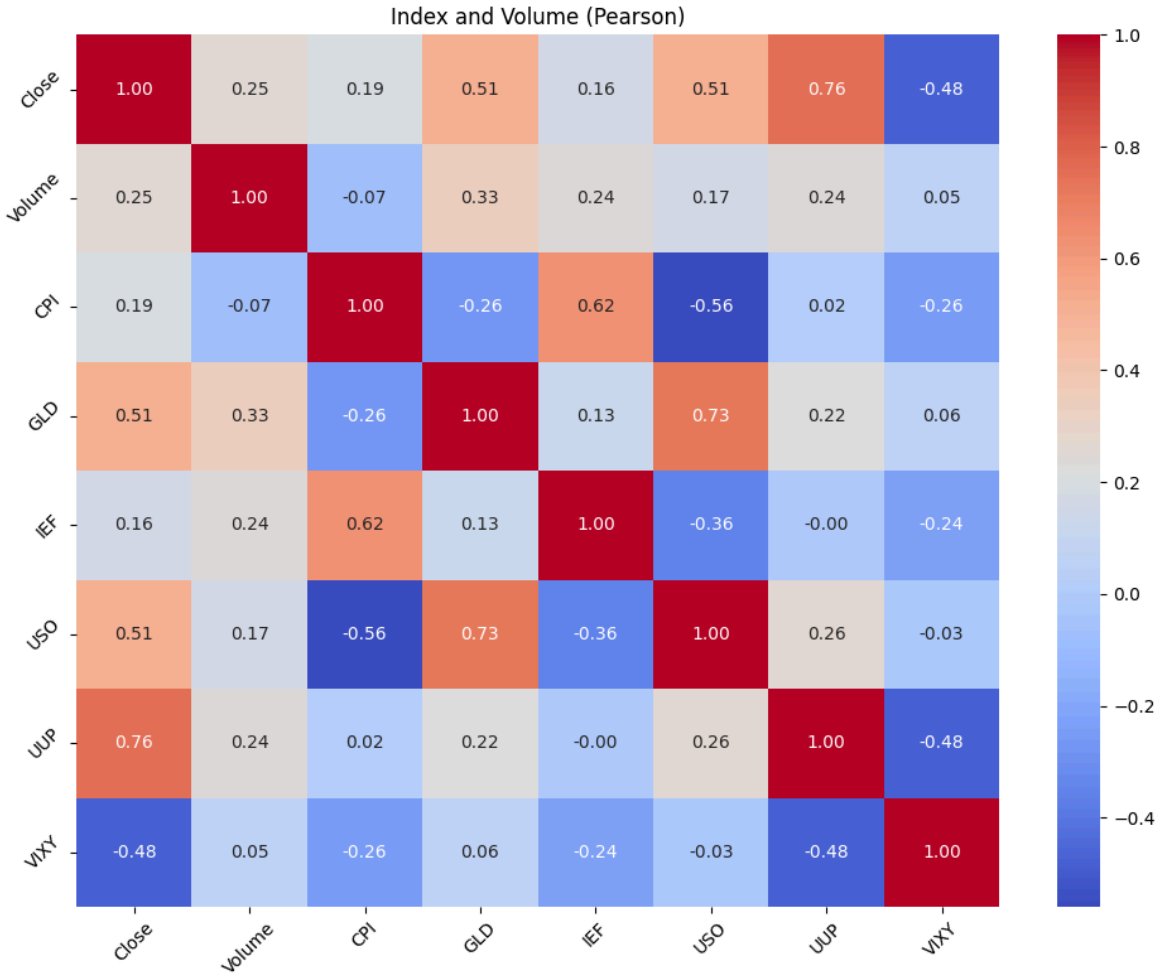

Figure 4.2.2. Pearson Correlation Matrix

- Close vs. VIX (-0.68): A strong negative correlation, as expected. This confirms the VIX as a "fear index": when the market (Close) drops, fear (VIX) rises.
- Close vs. DXY (0.90) / CPI_US (0.91): An extremely high positive correlation.
- Key Insight: This >0.90 correlation is clear evidence of Spurious

Conclusion: The correlation matrix confirms that running a regression on this raw, non-stationary data would be a mistake, as it is heavily biased by spurious correlations and multicollinearity.

CHAPTER 5. MULTIPLE LINEAR REGRESSION

5.1. Theoretical Basis

Multiple Linear Regression (MLR) is a statistical technique used to model the relationship between a dependent variable and multiple independent variables linearly:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

For time series data, a standard MLR often fails because consecutive observations are correlated. To incorporate temporal information while keeping the model linear, we introduce lag features of the dependent variable (e.g., Close_lag1, Close_lag2, Close_lag3). These allow the model to learn how past values of the target affect the current value.

All other variables (Open, High, Low, Volume, macro indicators) are included as independent predictors.

5.2. Model Implementation

- Target (y): Close
- Predictors (X): Open, High, Low, Volume, CPI, GLD, IEF, USO, UUP, VIXY, Close_lag1, Close_lag2, Close_lag3
- The model was trained using Linear Regression with a time-based train-test split (80%-20%).

5.3. Model Evaluation

Metric	**Value**
R^2	0.9962
RMSE	16.9717

Figure 5.3.1. Performance Metric

- The model explains 99.62% of the variance in Close prices. This is slightly lower than the naive linear regression without lag features, indicating more realistic predictive performance.
- RMSE ≈ 17: The average prediction error is around 17 units (e.g., dollars), which is reasonable for stock-level data.

5.4. Model Conclusion

Variable	Coefficient	P-value
Open	-0.6597	0.000
High	0.8511	0.000
Low	0.7980	0.000
Close_lag1	0.0402	0.008
Close_lag2	0.0215	0.007
Close_lag3	0.0069	0.253
Others	Near 0	> 0.05

Significant predictors: Open, High, Low, Close_lag1, and Close_lag2 — these variables meaningfully contribute to predicting Close.

Lag features:

- Close_lag1 and Close_lag2 have significant positive coefficients → past prices positively influence current Close.
- Close_lag3 is not significant, suggesting short-term lags capture most temporal dependency.

Other variables (macro indicators, Volume, VIXY): insignificant ($p > 0.05$) — they do not provide additional predictive power in this model.

Multicollinearity warning: Condition number ≈ 2.98e+11 — Open, High, Low, and lagged Close are highly correlated. Coefficients should be interpreted with caution.

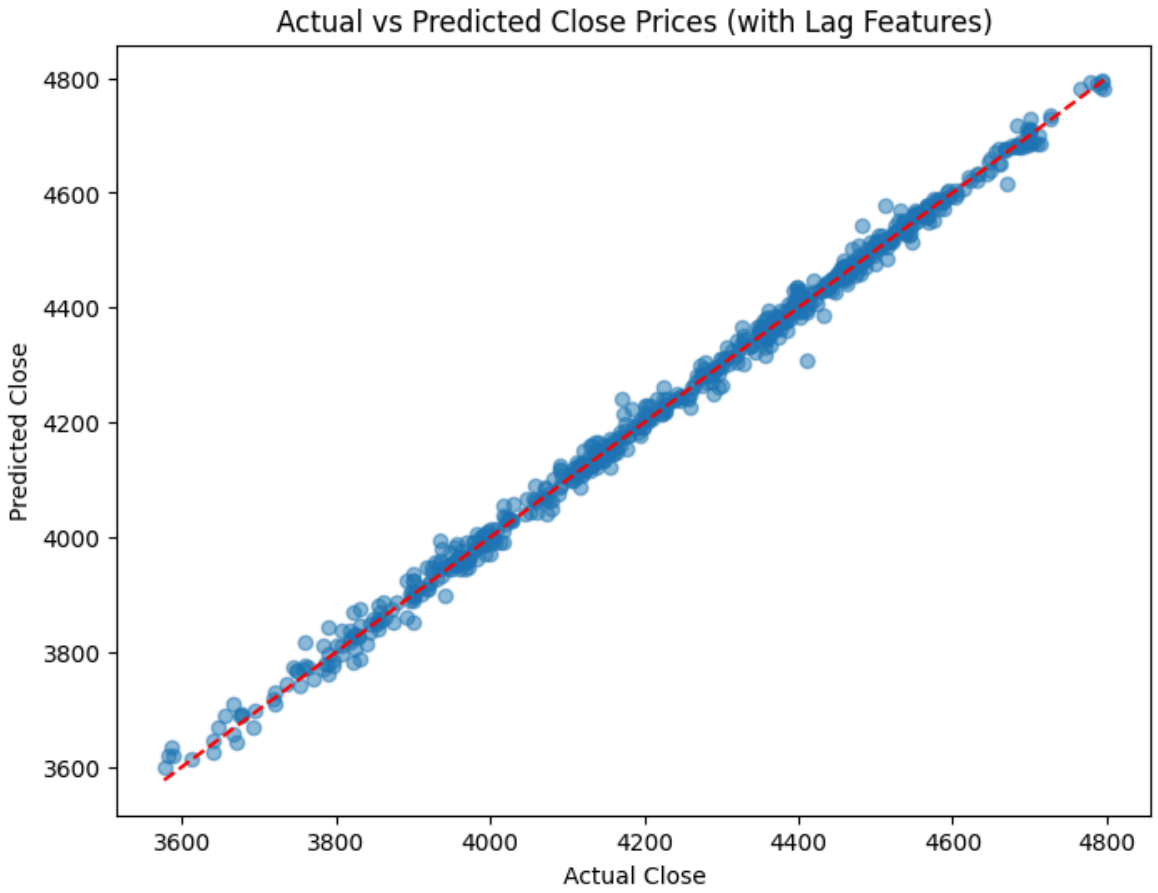

Figure 5.4.1. Actual vs Predicted Close Price (Lag feature added)